Understanding the Central Limit Theorem via an example.

Guillaume Dreyer

March 16, 2015

Overview.

We conduct a simple simulation and illustrate the meaning and main properties of the **Central Limit Theorem**. In particular, we show that the **distribution of the initial sample**, and the **distribution of the average of that sample**, are completely **unrelated**, which is the most powerful feature of the CLT.

Simulations.

Let's simulate 1000 averages of 40 **exponential distributions** with parameter rate **lambda equal to 0.2**.

```
head(averageExpo, 8)
```

```
##      value
## 1 5.461828
## 2 5.290509
## 3 5.046768
## 4 4.701764
## 5 4.880288
## 6 4.529606
## 7 5.344102
## 8 5.267107
```

The code (**Code 1**) for this simulation can be found in the **Appendix**. Briefly, **Code 1** does the following. It sets the seed for the 40000 simulations we want to generate. It stores the generated 40000 numbers in the data frame **simulations** that contains 40 columns and 1000 rows. It then takes the mean of each row of the latter, and stores the obtained values in the data frame **averageExpo**, that contains 1 column and 1000 rows.

Sample Mean versus Theoretical Mean.

1. Computations.

Recall that we are simulating the **distribution of the average of 40 exponential distributions**. The **sample mean** for the latter is obtained by taking the average of the column in the sample dataset **averageExpo**.

Regarding the **theoritical mean**: since the 40 exponential distributions all have a **mean** equal to **1/lambda** (with lambda = 0.2 in the present case), their average also has a **mean** equal to **1/lambda**.

```
sample_mean <- colMeans(averageExpo)
theoretical_mean <- 1/0.2       # the theoritical mean is 1/lambda
```

We obtain 5.0375393 for the **sample mean**, and 5 for the **theoretical mean**. One can see that these two values are quite close, as predicted by the **Law Of Large Numbers**.

2. Density of the average of 40 exponentials.

The histogram in **Fig. 1** of the **Appendix** displays the **density of the average of 40 exponentials**. The **red curve** interpolates the **underlying continuous density**. In addition, we added two vertical lines: the **red line** corresponds to the **sample mean** 5.0375393, and the **green one** to the **theoretical mean** 5. These two values being very close, the two lines naturally appear very near each other.

Sample Variance versus Theoretical Variance.

1. Computations.

As for the **sample mean**, the **sample variance** for our distribution is obtained by simply calculating the variance of the column in the sample dataset **averageExpo**.

Regarding the **theoretical variance**: recall that, if n independent random variables all have the same variance **var**, then the variance of their average is exactly **var/n**. In the present case, **n = 40** and **var = (1/.2)^2 = 25**. Hence

```
sample_variance <- var(averageExpo$value)
theoretical_variance <- 25/40
```

We obtain 0.6496971 for the **sample variance**, and 0.625 for the **theoretical variance**. Again, it is a consequence of the **Law Of Large Numbers** that these two values are very close.

2. Density distribution of the average of 40 exponentials.

The histogram in **Fig. 2** of the **Appendix** displays the **density of the average of 40 exponentials**. The **red curve** interpolates the **underlying continuous density**, and the vertical **red line** corresponds to the **sample mean** 0.6496971. We also added two other vertical lines: the **purple line**, whose distance from the **sample mean** 5.0375393 is the **standard error** 0.8060379; and the **blue line**, whose distance from the **sample mean** is the (theoretical) **standard deviation** 0.7905694.

Compare the density of the average of 40 exponentials to that of a single exponential.

In **Fig. 3** of the **Appendix**, the graph on the **left hand-side** displays the **density of the average of 40 exponentials**, whereas the one on the **right hand-side** displays **density of the exponential distribution**.

The left graph also shows (**green curve**) the **density of a normal distribution** with respective **mean** and **variance** equal to **5** and **25/40=0.625**. This graph perfectly illustrates the **Central Limit Theorem**: the distribution of the average of 40 independent identically distributed random variables (with mean **mu** and variance **sigma^2**) can be approximate by a normal distribution of mean **mu** and variance **sigma^2/40**. This is possible as the **size** of the sample (40 in the present case) is **"large enough"**.

Also, the right graph enables us to highlight **the most powerful feature of the CLT**. Evidently, by comparing the two graphs, the exponential distribution, and the average of 40 exponentials, look very different. In particular, one can see that **there is no correlation between the underlying distribution of the random variables occuring in the sample, and the distribution of the average of that sample**.

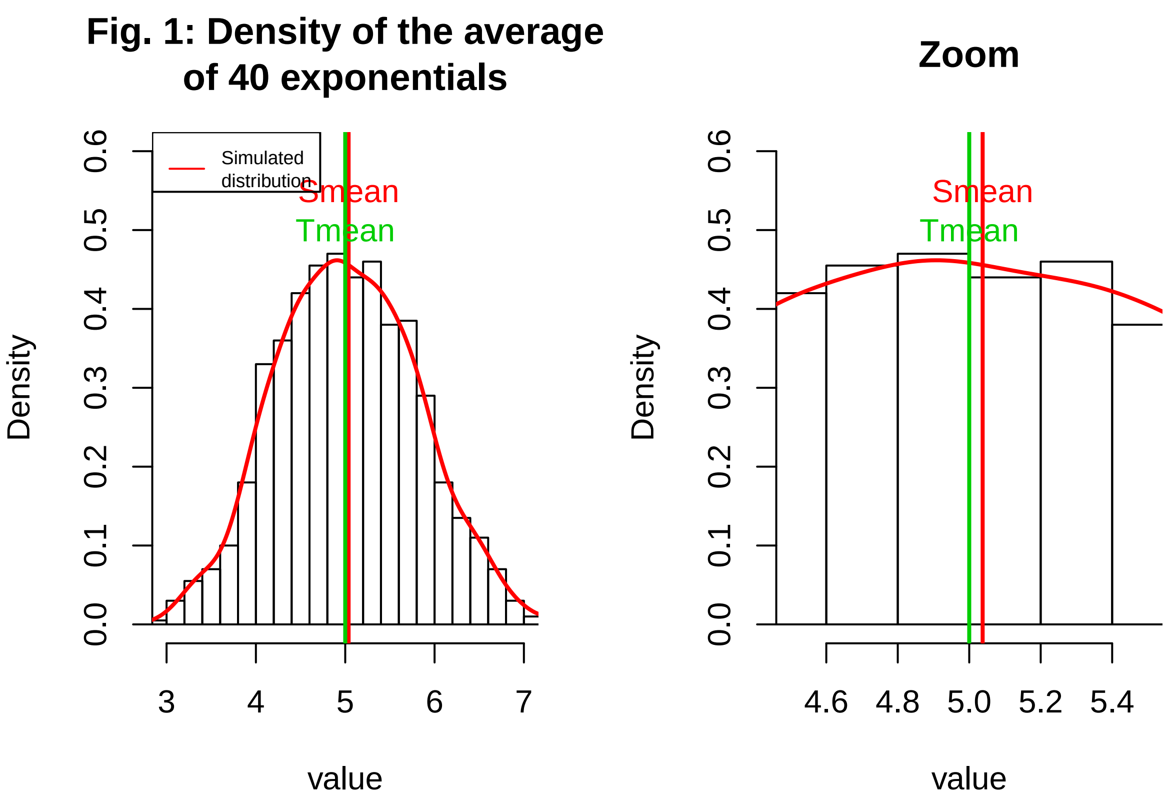
Appendix.

```
# Code 1.
nsim <- 1000
set.seed(40 * nsim)      # set the seed for the 40000-size sample
                         # prior to the simulation

simulations <- as.data.frame(matrix(data = rexp(n = 40 * nsim, rate = 0.2),nrow = 1000))
                         # generate 40000 simulations of an
                         # exponential distribution, and store
                         # them in the data frame "simulations"

averageExpo <- data.frame("value" = rowMeans(simulations))
                         # take the mean of the rows of the data
                         # frame "simulations", and store the values
                         # in the data frame "averageExpo"
```

```
graph1();
```




```
graph2();
```

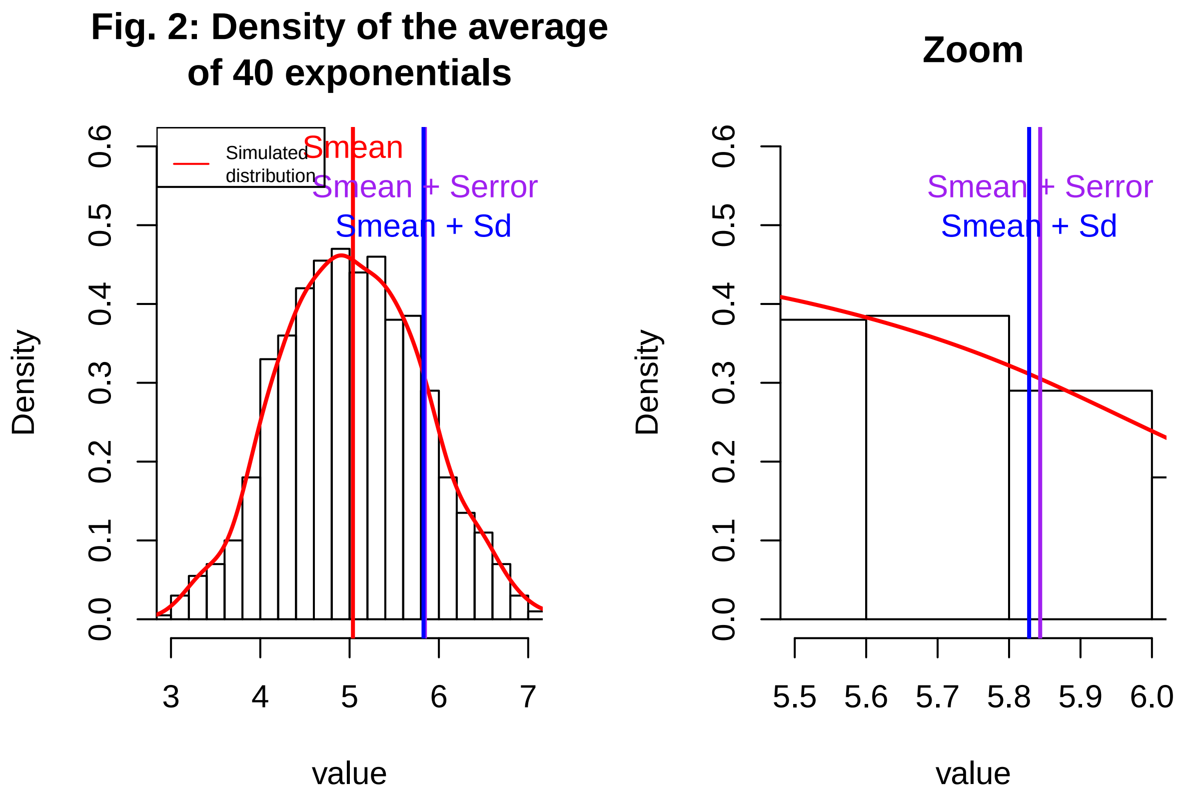


Fig. 2: Density of the average
of 40 exponentials
Simulated
distribution
Smean
Smean + Serror
Smean + Sd
Density
value
Zoom
Smean + Serror
Smean + Sd
Density
value

```
graph3();
```

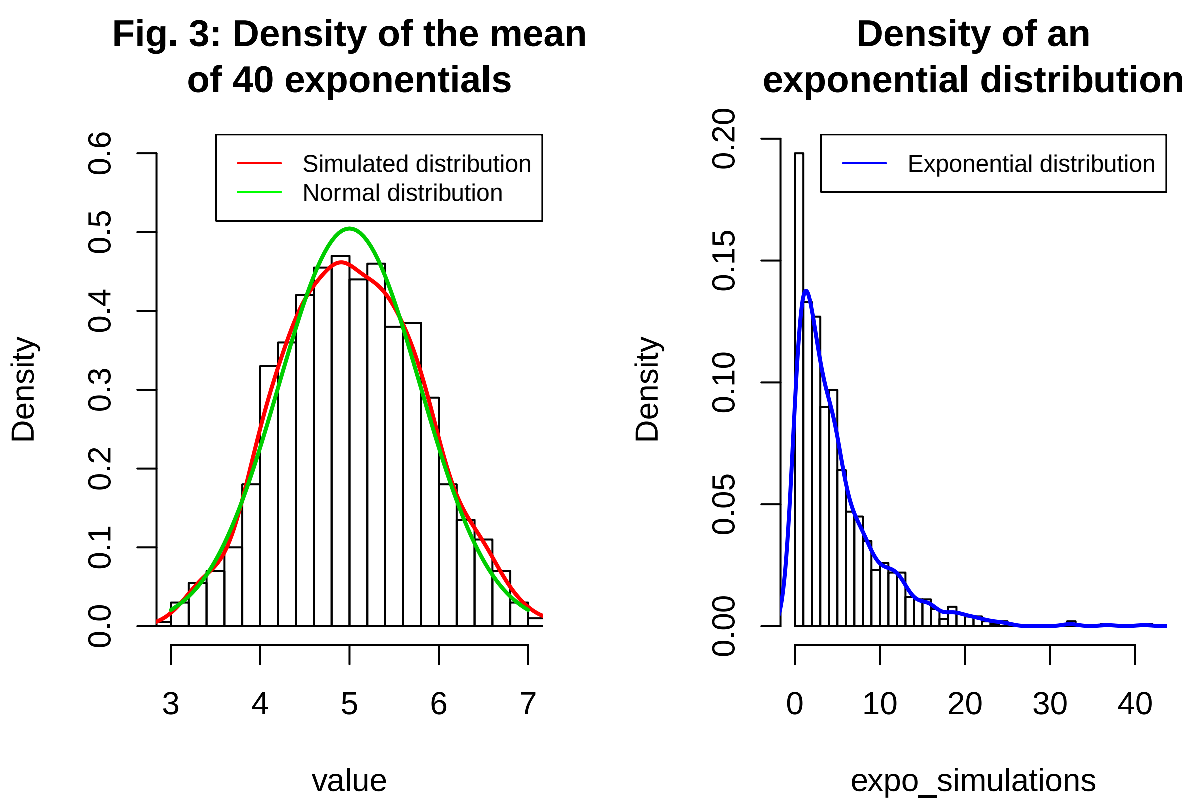


Fig. 3: Density of the mean
of 40 exponentials
Simulated distribution
Normal distribution
Density
value
Density of an
exponential distribution
Exponential distribution
Density
expo_simulations

```
graph1 <- function(){
        par(mfrow=c(1,2))
        hist(averageExpo$value,
             main = "Fig. 1: Density of the average\nof 40 exponentials",
             freq = FALSE, breaks = 30,
             xlim=c(3, 7), ylim=c(0, 0.6), xlab="value")
        lines(density(averageExpo$value), lwd = 2,col = "red")
        abline(v = sample_mean, col = 'red', lwd = 2)
        abline(v = theoretical_mean, col = 'green3', lwd = 2)
        text(sample_mean, 0.55,"Smean",col = 'red')
        text(theoretical_mean, 0.50,"Tmean",col = 'green3')
        legend("topleft", legend = c('Simulated\ndistribution'),
               lwd =c(1), col=c('red'),cex=0.6)

        hist(averageExpo$value, main="Zoom",
             freq = FALSE, breaks = 30, xlim=c(4.5, 5.5),
             ylim=c(0, 0.6), xlab="value")
        lines(density(averageExpo$value),lwd = 2,col = "red")
        abline(v = sample_mean, col = 'red', lwd = 2)
        abline(v = theoretical_mean, col = 'green3', lwd = 2)
        text(sample_mean, 0.55,"Smean",col = 'red')
        text(theoretical_mean, 0.50,"Tmean",col = 'green3')
        }
```

```
graph2 <- function(){
        par(mfrow=c(1,2))
        hist(averageExpo$value,freq = F, breaks = 30,
             main = "Fig. 2: Density of the average\nof 40 exponentials",
             xlab="value",xlim=c(3, 7), ylim=c(0, 0.6))
        lines(density(averageExpo$value), lwd = 2,col = "red")
        abline(v = sample_mean + sqrt(sample_variance),
               col = 'purple', lwd = 2)
        abline(v = sample_mean + sqrt(theoretical_variance),
               col = 'blue', lwd = 2)
        abline(v = sample_mean, col = 'red', lwd = 2)
        text(sample_mean, 0.6,"Smean",col = 'red')
        text(I(sample_mean + sqrt(sample_variance)), 0.55,
             "Smean + Serror",col = 'purple')
        text(I(sample_mean + sqrt(theoretical_variance)), 0.50,
             "Smean + Sd",col = 'blue')
        legend("topleft", legend = c('Simulated\ndistribution'),
               lwd =c(1), col=c('red'),cex=0.6)

        hist(averageExpo$value,freq = F, breaks = 30,
             xlim = c(5.5, 6), ylim=c(0, 0.6),
             main = "Zoom",xlab ="value")
        lines(density(averageExpo$value), lwd = 2,col = "red")
        abline(v = sample_mean + sqrt(sample_variance),
               col = 'purple', lwd = 2)
        abline(v = sample_mean + sqrt(theoretical_variance),
               col = 'blue', lwd = 2)
        abline(v = sample_mean, col = 'red', lwd = 2)
        text(I(sample_mean + sqrt(sample_variance)), 0.55,
```

```
             "Smean + Serror",col = 'purple')
        text(I(sample_mean + sqrt(theoretical_variance)), 0.50,
             "Smean + Sd",col = 'blue')
        }
```

```
graph3 <- function(){
        par(mfrow=c(1,2))
        hist(averageExpo$value,freq = F, breaks = 30,
             main = "Fig. 3: Density of the mean\nof 40 exponentials",
             xlab="value", xlim=c(3, 7), ylim=c(0, 0.6))
        lines(density(averageExpo$value), lwd = 2,col = "red")
        curve(dnorm(x, mean=5, sd=sqrt(25/40)), add=TRUE,
              col="green3",lwd=2)
        legend("topright",
               legend = c('Simulated distribution',
                          'Normal distribution'),
               lwd =c(1, 1), col=c('red', 'green'),cex=0.75)

        set.seed(1000)
        expo_simulations <- rexp(1000, rate = 0.2)
        hist(expo_simulations,freq = F, breaks = 30,
             main = "Density of an\nexponential distribution")
        lines(density(expo_simulations), lwd = 2,col = "blue")
        legend("topright",
               legend = c('Exponential distribution'),
               lwd =c(1), col=c('blue'),cex=0.75)
        }
```